MCC Data Documentation and Anonymization Guidelines
June 24, 2015

Objectives

Millennium Challenge Corporation (MCC) is committed to publicly sharing microdata generated in the design, implementation, and evaluation of its compacts and threshold programs. For public release of this data, MCC considers three objectives:

1. *Maximizing replicability*. To enable any stakeholder, researcher, or agency to understand the source data and analysis behind MCC evaluations and investments.
2. *Maximizing usability*. Recognizing the value of data generated through MCC projects and investments, public access to MCC-financed data can stimulate a wide range of policy relevant research, maximizing the benefits of MCC's investments in large-scale data collection efforts in developing countries.
3. *Ensuring confidentiality of respondents*. The previous two objectives must be balanced with obligations to protect the confidentiality of survey respondents who are crucial to the production of microdata. There are two main forms of risk to the respondents: (i) risk of confidentiality of participation in the survey, and (ii) risk of loss of confidentiality of PII and other sensitive data. The informed consent should set forth the level of confidentiality promised to respondents.

These guidelines lay out the basic requirements for ensuring MCC-financed microdata is stored, transferred, and publicly accessible by two distinct audiences: (i) MCC staff and contractors, for internal analysis and decision-making, and (ii) public users, including academic researchers, policy analysts, and peer institutions.

Following these guidelines is requested for MCC staff, survey firms and independent evaluators responsible for household, firm, and enterprise survey data collected for independent evaluations of Compact activities. In many instances, data collection is contracted by the country MCA unit and funded from the country compact budget. Such data is considered to be funded by MCC and is under the purview of these guidelines. In other instances, MCAs and/or MCC may collaborate with other organizations engaged in survey data collection. In these instances, MCAs and/or MCC may fund only a portion of a dataset (such as an additional module or over-sampling in targeted regions). These datasets should also comply with these guidelines, with the exception of certain circumstances where local regulations or prior agreements with national statistical agencies restrict the dissemination of the data.

For other data collection activities, stakeholders should be aware of these guidelines and follow them to the extent possible.

Storage and Transfer

Regardless of the data handler (MCC, MCA, survey firm, evaluator, other), the following should be kept in mind:

1. Once data collection ends, there must be specific practices in place to protect confidentiality of the data during storage, regardless of the data form (paper-based or

electronic). This includes actions by any data handlers, such as: using lock and key cabinets for paper files; encrypting data files; employing password protection on data systems and data encryption; and requiring relevant stakeholders to sign non-disclosure agreements.

2. When sharing data files, data handlers should use a secure file transfer system.

Documentation

In order to facilitate access to and usability of data, *all datasets delivered to MCC must be accompanied with completed documentation, in the form of standardized metadata.* Based on the International Household Survey Network's (IHSN) Metadata Editor[1], the MCC has created an MCC Evaluation Metadata Template[2] (a Nesstar file template[3] is included as Annex I) that specifies the required metadata elements for documentation purposes of evaluation related data. This information is what is presented in the MCC Evaluation Catalog entry for that evaluation. These elements are designed to be compliant with the international Data Documentation Initiative (DDI)[4] and Dublin Core Metadata Initiative (DCMI)[5] standards, enabling compatibility with various data archiving systems. The Metadata can be created as soon as the Evaluation Design Report is approved and ready for public release.

The Metadata can be updated/revised as necessary once completed for the evaluation. In addition to completing the Metadata Template, when submitting a data package consultants should submit all necessary materials as summarized in Table 1 below.

In addition to summarizing the core documents, Table 1 also provides the recommended format for each core document. All text documents – including reports, manuals, questionnaires, and codebooks – must be available in both their *original editable format* (Word, Excel, etc.) and in searchable Portable Document Format (PDF).

For quantitative data, the preferred file format is STATA; however, other common statistical formats (SPSS, SAS, etc.) are acceptable as per agreement with MCA/MCC. Command file and data file formats should be the same. The data files must include variable names, labels, response value labels or formats, and clearly defined associations between survey questions and the variables containing the responses.

When the data collection occurs in a language other than the four commonly used languages (English, French, Portuguese and Spanish), the information in the data file should include common-language versions. When the resources required for translation are considered reasonable, the focus of the translation should be on any variables that are used in the primary

[1] http://www.ihsn.org/home/index.php?q=tools/toolkit

[2] MCC is working to develop a Nesstar Metadata Template to mirror the excel file created for the MCC Evaluation Catalog. When the Nesstar Metadata Template is available, these guidelines will be updated to reflect requirements for submission of documents in the Nesstar file.

[3] To use the template, open Nesstar, select 'Documentation' -> 'Templates' -> 'Import' -> select template. Ensure the MCC Template 1.3 is checked as the default template before starting.

[4] http://www.ddialliance.org/

[5] http://dublincore.org/

impact estimation (i.e. providing English variable names and/or labels). Other elements, such as questionnaires and enumerator and trainer manuals, should be available in all languages used in fielded questionnaires and in a common-language translation (preferably English).

Table 1: Summary of Study Documentation

Element	Requested Format	Description
"Readme" File	Word or PDF	A "Readme" file briefly outlining the contents of the survey package, listing all included files and their formats and purposes. MCC has a specific template for this file.
Enumerator and trainer manuals	Word, searchable PDF	Guides for survey enumerators, supervisors, and trainers, as well as manuals used to train each of these individuals
Questionnaires	Word, Excel, or other as appropriate	Survey instruments/questionnaires
Original, Raw Data	STATA (or other format agreed with MCC)	This is the complete data file(s) submitted by the survey firm with appropriate and logical variable names and labels, and including any necessary personally identifiable information of survey respondents. Data files should be submitted in Stata version 11.0 or higher.
Public Use Data	STATA (or other format agreed with MCC)	This is anonymized data following the Anonymization Worksheet (section 3). Data files should be submitted in Stata version 11.0 or higher.
Public Use Data Codebook	PDF	Codebook of public use data files
Analysis Data	STATA (or other format agreed with MCC)	MCC is specifically committed to enabling the re-creation of results produced by independent evaluators. In order to do so, consultants must either provide full STATA do files used to produce analysis files from public use files, or submit anonymized analysis files as well. Submission of anonymized analysis files only is insufficient considering the potential for errors in coding and variable construction. Therefore, sufficient documentation is required in order to define how analysis files are produced from public use data.
Evaluation Design Concept Note, Baseline Report and Impact Report	Word or searchable PDF	These documents (deliverables required by the contract terms of reference) provide additional useful design and analytical information for users of the data – particularly if this information cannot be reasonably included in the metadata. Evaluators should ensure that all public use documents/reports have been reviewed and edited to remove any references, such as geographic locations, that may threaten or undo anonymization efforts. In these cases, the evaluator should also provide internal use only documents/reports that include all removed information.
Analysis Programs & Command Files	STATA do files (or other format agreed with MCC)	Programs, command files, and/or ".do" files used in the preparation of the data and the analyses presented in the Final Report. These programs can be provided in any format using any statistical package or software, although Stata is preferred. They should include data merging, imputing, and other preparation work, key summary tabulations and estimations included formally in the Report. As applicable, they should be clearly labeled so as to correspond with tables or sections in the Report.
Other study documents	As appropriate	This includes Institutional Review Board approvals, informed consent statements, other documentation for the study
Data Dissemination Statement	Searchable PDF	Some datasets are subject to prior agreements or local legal restrictions preventing public dissemination. This includes agreements with academic researchers delaying data release until after an agreed-upon date. If any of these situations apply, please include a brief statement clarifying why data cannot be published or when the data may be published.

Anonymization

MCC seriously considers the ethical and privacy implications of research involving human subjects. As such, MCC respects, and where feasible, follows the Common Federal Policy for the Protection of Human Subjects, also known as the "Common Rule," in addition to applicable privacy laws and regulations.

To protect the rights and privacy of individual respondents to MCC-funded surveys, MCC requires public use data files to be free of personal or geographic identifiers that would permit unassisted identification of individual respondents or their household members, and to exclude variables that introduce reasonable risks of deductive disclosure of the identity of individual subjects.

While there are established procedures for de-identifying quantitative data sets, removing PII from qualitative data sets is more difficult. For this reason, when data sets cannot be disseminated without the risk of identification of respondents, MCC will only release questionnaires and codebooks.

These standards are drawn from best practices outlined by the International Household Survey Network[6], as well as recommendations made by the Office of Management and Budget, the US Census Bureau, and USAID's Demographic and Health Surveys.

Important note: Recognizing that there is a fine balance between anonymization and data usability, MCC is receptive to varying levels of anonymization. Although all public use data should be fully anonymized to the extent possible, MCC may also choose to release more complete data on a restricted-access basis, which will require direct requests by the end-user and a full assessment by MCC of the end-user's request, in accordance with necessary Institutional Review Board approval.

Consultants should complete the Anonymization Worksheet (Annex II) and include it in the delivery of public use, anonymized data. Unless marked as "*Optional*," the standards in the worksheet represent the MCC's minimum standards and must be adhered to for all datasets intended for public use.

Consultant should expect to submit the following anonymization package to the MCC Disclosure Review Board (DRB) for review of any public use data:

- Completed Anonymization Worksheet
- Metadata
- Codebooks of public use data
- Public use data files
- Informed consent statement
- Questionnaires

[6] http://www.ihsn.org/home/node/137; http://www.ihsn.org/HOME/node/138

Consultant may determine it is more appropriate to submit a proposal for the anonymization efforts for review and discussion with the DRB prior to official submission. In these cases, the consultant should submit:

- Completed Anonymization Worksheet
- Metadata
- Informed Consent statement
- Questionnaire

Local Laws

In addition to the guidelines found here, data handlers should comply with all relevant local laws, including those that may override these guidelines or prevent dissemination. Where local law prevents compliance with these guidelines, please include a brief explanation.

Review of these Guidelines

These guidelines may be reviewed and updated. Consultants should work off the most recent, approved version of these guidelines.